

Diabetes mellitus diagnosis method based random forest with bat algorithm

Syaiful Anam¹, Fidia Deny Tisna Amijaya², Satrio Hadi Wijoyo³, Dian Eka Ratnawati⁴,

Cynthia Ayu Dwi Lestari¹, Muhammin Ilyas¹

¹Department of Mathematics, Faculty of Mathematics of Natural Science, Brawijaya University, Malang, Indonesia

²Department of Mathematics, Faculty of Mathematics of Natural Science, Mulawarman University, Samarinda, Indonesia

³Department of Informatics Engineering, Faculty of Computer Science, Brawijaya University, Malang, Indonesia

⁴Department of Information System, Faculty of Computer Science, Brawijaya University, Malang, Indonesia

Article Info

Article history:

Received Feb 1, 2024

Revised Nov 5, 2024

Accepted Nov 14, 2024

Keywords:

Bat algorithm

Diabetes mellitus

Diagnosis

Hyperparameter optimization

Random forest

ABSTRACT

Diabetes mellitus (DM) is a very dangerous disease and can cause various problems. Early diagnosis of DM is essential to avoid severe effects and complications. An affordable DM diagnosis method can be developed by applying machine learning. Random forest (RF) is a machine learning technique that is applied to develop a DM diagnosis method. However, the optimization of RF hyperparameters determines the performance of RF approach. Swarm intelligence (SI) could be used to solve the hyperparameter optimization problem on RF. It is robust and simple to be applied and doesn't require derivatives. Bat algorithm (BA) is one of SI techniques that gives a balance between exploration and exploitation to find a global optimal solution. This article proposes developing an RF-BA-based technique for diagnosing DM. The results of the experiment demonstrate that RF-BA can diagnose DM more accurately than conventional RF. RF-BA has higher performance compared to RF-particle swarm optimization (PSO) in terms of computational time. The RF-BA also are able to solve the overfitting problem in the conventional RF. In the future, the proposed method has a high chance of being implemented for helping people with early DM diagnosis with high accuracy, low cost, and high-speed process.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Syaiful Anam

Department of Mathematics, Faculty of Mathematics of Natural Science, Brawijaya University

Veteran Street, Malang, Indonesia

Email: syaiful@ub.ac.id

1. INTRODUCTION

A severe metabolism that elevates blood sugar levels is a hallmark of diabetes mellitus (DM) [1]. Indonesia is rated seventh out of ten countries in the world for the overall number of DM patients. Ten point eight million people in Indonesia will have DM in 2020, representing 6.2 percent of the country's total patient population [2]. Various complications are brought on by DM [3], [4]. Persons who have type 1 or type 2 diabetes frequently experience complications and they also dramatically raise mortality as well as morbidity [5]–[7]. There are two main categories of complications associated with diabetes which are microvascular and macrovascular. The microvascular complications have a significantly greater frequency than the macrovascular complications [8]. Microvascular problems include retinopathy, neuropathy, and nephropathy, whereas macrovascular complications include peripheral artery disease, stroke, and cardiovascular disease [3], [9], [10]. With 236 thousand DM-related deaths in 2021, Indonesia ranks as having the sixth-highest DM mortality rate, according to the International Diabetes Federation (IDF) [11].

Early identification of DM is crucial to prevent its severity symptoms and complications. Diagnosis of DM is typically performed by a highly skilled physician and needs high expense [12], [13]. The results of the diagnosis by the physicians are often biased amongst specialists [14]. A reliable and reasonably priced method for diagnosing DM is possible to be done by employing a classification technique. One of the popular and robust classification methods is machine learning. Machine learning method has many advantages over traditional methods. Their advantages have been extensively studied and documented in various fields. Machine learning methods have demonstrated several benefits over traditional approaches, including improved performance, enhanced predictive capabilities, and the ability to handle complex data structures [15]. The machine learning also no need strong assumptions about the type of error distribution, much more flexible and do not require any a priori assumptions [16]. Machine learning has been applied successfully in many fields, such as vehicular networks [17], [18], medical diagnosis [19], [20], speech recognition [21], computational imaging [15], medical healthcare [22], signal processing [23], and autonomous driving [24]. The machine learning technique known as random forest (RF) has numerous benefits, such as the capacity to manage big datasets with high dimensionality, ease of use, resistance to outliers and noisy data, easy parallelization, good avoidance of overfitting, rapid processing, excellent precision, robustness, and a wide variety of variables [25]–[28].

The RF approach has been utilized to some applications, including predicting consumer churn [29]–[31], detection of heart disease [32], insurance acceptance prediction [33], identifying fraud [34], loan forecasting [27], and breast cancer detection [35]. However, the capability of the RF technique is greatly impacted by the choice of hyperparameters. When the hyperparameters are selected incorrectly, the loss of function cannot be efficiently reduced, which leads to imprecise findings from the RF approach. Therefore, the right RF hyperparameter must be chosen or optimized to maximize the efficacy of the RF approach. Several investigations have shown that hyperparameter adjustment significantly improves RF performance [36]. In study by Zhu *et al.* [36], the grid search is used to select the RF hyperparameters. However, this method needs a high computational cost, since all combinations of RF hyperparameters have to be tried to find the best hyperparameters. The selection of hyperparameters of the RF problem can be represented in the optimization formulation. For this reason, RF can be combined with global optimization methods, such as the swarm intelligence (SI) technique, to solve the issue of choosing hyperparameters in the RF method. The RF hyperparameters optimization by using SI doesn't have to try all combinations of RF hyperparameters in the search domain, which means that the RF hyperparameters optimization by using SI may result in a shorter computational time than the grid search method.

In addition, the SI algorithm offers a number of benefits. The SI approach has ability to search a global optimum in multimodal functions, is resilient, easy to implement, and doesn't need derivative [37]. Numerous SI techniques have been put forth. Bat algorithm (BA), artificial bee colony (ABC) algorithm, particle swarm optimization (PSO), and firefly algorithm (FA) are some SI examples. According to certain research, PSO and BA provide advantages in the balance between exploration and exploitation. BA has a number of benefits, including quick convergence and the requirement for few parameters [38]. Additionally, one study demonstrates that the convergence of BA is better than the genetic algorithm (GA) and PSO [39]. For the diagnosis of DM, the K-means algorithm, which BA optimized, has been used [40]. Research indicates that the K-means algorithm's performance can be considerably enhanced by the BA approaches; nevertheless, other studies' findings indicate that the RF method outperforms the K-means method [41]. BA also has been applied successfully to many fields such as transport network design problem [42], job scheduling problem [43], image enhancement [44], and disease classification [45].

Based on the problem that has been described, this article suggests developing an RF with a hyperparameter Bat algorithm optimizer (RF-BA) for DM diagnosis. BA is employed to optimize the RF hyperparameters. This article has contributed to creating a DM diagnosis method with high accuracy and acceptable computational time. This article also has a contribution to selecting the RF hyperparameters for increasing the performance of RF by utilizing BA with shorter computation time than the computational time of the previous method in [36]. The suggested approach is assessed using a number of performance criteria, including computation time, f1 score, accuracy, recall, and precision. The suggested approach is contrasted with RF-PSO and traditional RF.

2. METHOD

The development of the proposed method will be covered in this part. There are multiple steps in the process, including: i) gather data, ii) pre-processing data, iii) develop RF-BA method, iv) set parameters of the proposed method, v) assess the proposed method, and vi) draw conclusion.

2.1. Dataset

The dataset came from the kaggle.com. The RF-BA approach for diagnosing DM is developed using certain features. The features used in the classification model for the diagnosis of DM include high blood pressure (*hbp*), high cholesterol (*hc*), no cholesterol check in five years (*chol*), body mass index (*bmi*), smoker (*smk*), stroke (*str*), disease or heart attack (*ha*), physical activity (*pa*), fruits (*frt*), vegetables (*vgt*), heavy drinkers (*hd*), need to see a doctor (*nsd*), general health (*gh*), mental health (*mh*), physical health (*ph*), difficult walk (*dw*), sex (*sx*), age (*ag*), education (*ed*), income (*inc*), and diabetes (*D*). Next, it will be discussed an explanation of each variable. *hpb* indicates patient with hypertension and *hc* represents for patient cholesterol level. The patient has smoked at least 100 cigarettes during their lifetime, according to the *smk*. *str* defines the patient with stroke. Patients with myocardial infarction (MI) or coronary heart disease (CHD) are defined by *ha*. *pa* represents the activity of patient in past 30 days, not including job. *frt* are defined as those that are consumed at least once every day and *vgt* are defined as those that are consumed at least once every day. While, adult men who consume more than 14 drinks per week and adult women who consume more than seven drinks per week are considered *hd*. *nsd* denotes a period within the previous 12 months when a patient needed to see a doctor but was unable to do so due to financial constraints. *gh* is measured on a scale of 1 for excellent, 2 for very good, 3 for good, 4 for fair, and 5 for poor and *mh* includes stress, depression, and emotional issues, as well as the number of days in the previous 30 days that were not good for mental health. Physical disease and injury are included in *ph*, as is the number of days in the last 30 days where physical health was poor or nonexistent. The *dw* symbolizes the extreme difficulty of ascending stairs or walking. The 13-level age category is determined by *ag*. On a ranking system of 1 to 6, *ed* represents educational level. Value of 1 represents only attending preschool or never attending school, 2 represents completing elementary school grades 1 through 8, 3 indicates some of the high school grades 9 through 11, 4 indicates grade 12 or high school graduate (GED), 5 represents one to three years of college, and 6 represents four years or more of college (college graduate). According to the income scale, a value of 1 denotes less than \$10,000, a value of 5 denotes under than \$35,000, and a value of 8 denotes greater than \$75,000. *D* represents the DM state.

2.2. Data pre-processing

Preparing raw data into a useful format is the aim of data pre-processing. Several data pre-processing methods were used in this work, such as data transformation, dealing with missing values, and missing value inspection. Inadequate handling of these missing values will likely make it difficult to draw a trustworthy conclusion. The data in this study is transformed using min-max normalization. Data transformation's main objective is to change the scale of measurement of the raw data into a different format so that it can be processed effectively and satisfy the specifications of the selected processing method.

2.3. Developing the algorithm of the DM diagnosis method based RF-BA

The development of the algorithm for the RF-BA based on the DM diagnosis approach will be discussed in this part. The RF hyperparameters are optimized to create the RF-BA based on the DM diagnosis method. RF is devised by Breiman and Cutler. First, the way of RF works will be described here. The workings of the RF to solve classification problems can be seen in Figure 1. The RF is made up of multiple decision trees that were constructed with random vectors. The RF algorithm can be expressed simply as follows: let us assume that the training data set comprises *p* predictor variables and has a size of *n* observations.

The steps involved in RF estimation and preparation are [46]: i) the bootstrap stage is to draw random samples of size *n* from *h* training data; ii) utilizing a bootstrap dataset, the tree is constructed until it achieves its optimum size (without pruning) in the random sub-setting step. The sorter is selected at each node by selecting *m* predictive variables at random, where $m < p$. The best sorter is then selected based on the *m* predictor variables; iii) to create a forest made up of *k* RF, repeat the procedure 1-3 *k* times; iv) voting stages: the voting stage is carried out for each prediction result, for classification problems the mode will be used, and for regression problems the mean will be used; and v) the final step is that the algorithm will select the most frequently selected prediction results as the final prediction.

In this work, the settings in the RF are optimized using BA. RF performance is enhanced by precise settings. The flowchart for the DM diagnosis method based RF-BA is displayed in Figure 2. There are several phases in the suggested method. In the initial step, the dataset is entered and divided into data for training and testing. Additionally, the parameters of BA also should be entered, such as number of bat (*Nbat*), the maximum number iteration (*t_{max}*), loudness of bat (*A*), *pulse rate* (*r₀*), α , γ , maximum frequency (*f_{max}*) and minimum frequency (*f_{min}*).

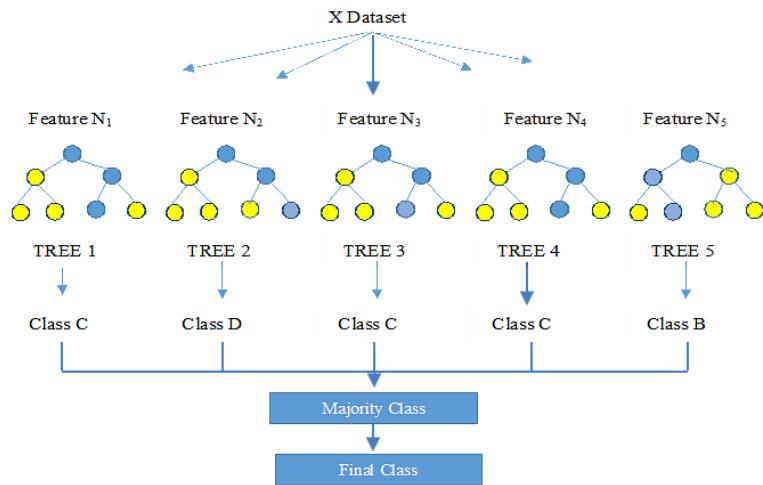


Figure 1. Illustration of RF

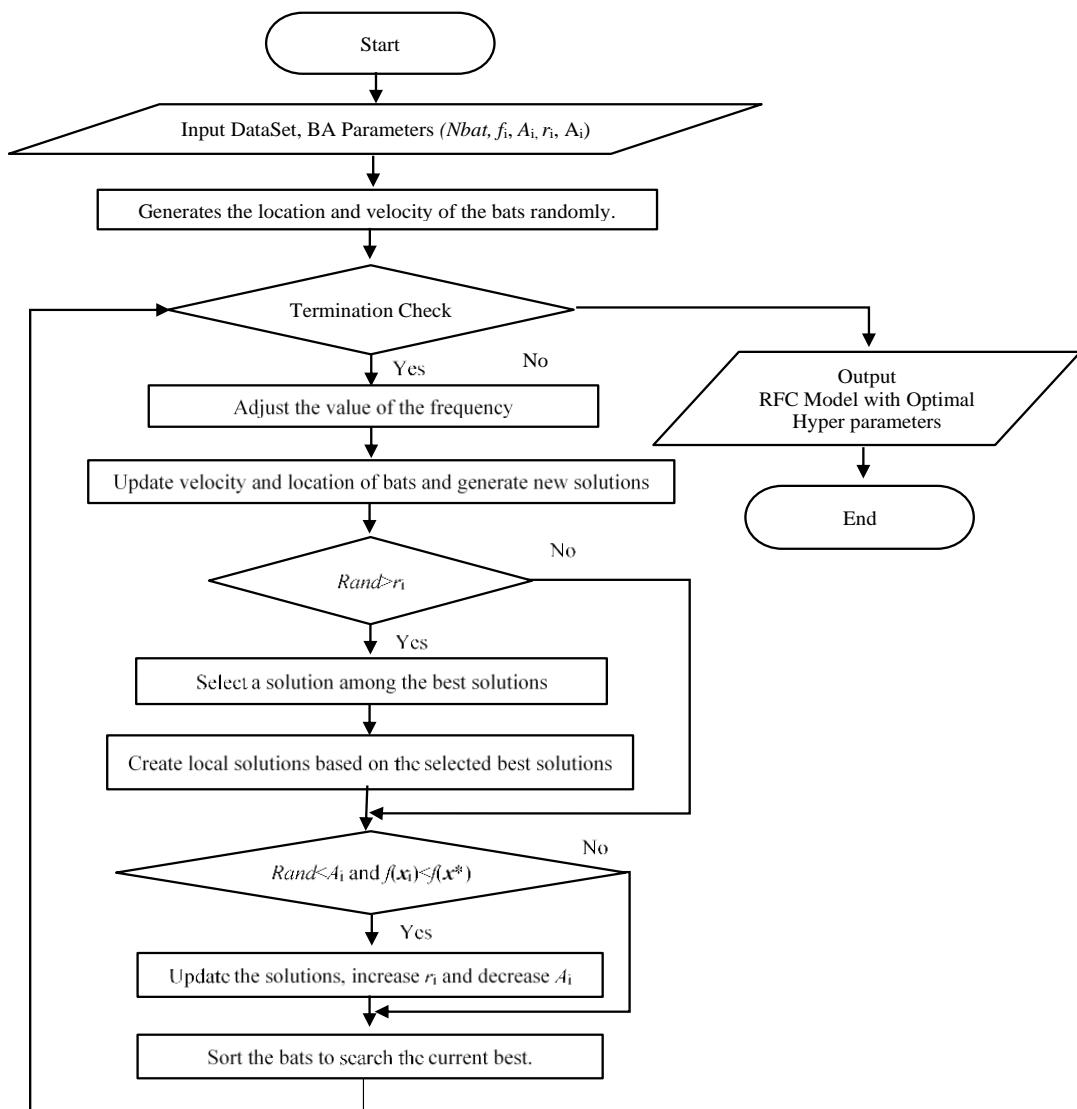


Figure 2. Flowchart of RF-BA

As a result, bat positions and velocities are determined at random. Each bat's position determines the RF hyperparameter. The RF parameters are minimal sample split (*max_depth*), maximum depth (*max_feat*), minimum sample split leaf (*max_Sam_Split*), and the number of estimators (*n_est*). The number of trees in RF defines the number of estimators. The total number of nodes on the longest path from the root node to the leaf node is known as the maximum depth. The bare minimum of samples needed to split an internal node is known as the minimal samples split. The minimal number of samples needed at a leaf node is known as the minimum sample split leaf. The bat's position is specified in (1).

$$x_i = (x_{i,1}, x_{i,2}, \dots, x_{i,4}), i = 1, \dots, N_{bat} \quad (1)$$

The number of particles is defined by *Nbat*. The *i*-th bat, denoted by x_i , is the RF parameter candidate. The number of estimators is represented by $x_{i,1}$, the maximum depth by $x_{i,2}$, the minimal sample split by $x_{i,3}$, and the minimum sample split leaf by $x_{i,4}$. Every parameter has a range of differences. The following section will provide a description of the range of parameters. The following stage is adjusting each bat's frequency value using (2).

$$f_i = f_{min} + (f_{max} - f_{min})\beta \quad (2)$$

Following each bat's frequency adjustment, the bats' position and velocity are updated, and new solutions are produced using (3), (4), and (5), respectively.

$$v_i^{t+1} = v_i^t + (x_i^t - x_*)f_i \quad (3)$$

$$x_i^{t+1} = x_i^t + v_i^{t+1} \quad (4)$$

A number at random with a uniform distribution $U(0,1)$ is utilized to produce the optimal solutions. Using the fitness function in Pseudocode 1, the optimal solution determines each bat's fitness value. If ($rand < r_i$), then (5) generates the local solutions at random.

$$x_{new} = x_{old} + \sigma \epsilon_t A^{(t)} \quad (5)$$

$A^{(t)}$ is the mean of bat loudness over time t , σ is the scaling factor, and ϵ_t displays random values obtained from a distribution that is normal that has the Gaussian shape $N(0,1)$. σ is equal to 0.01 which could possibly be used for practicality.

Pseudocode 1. Fitness function

```
Function fitness (x, X_training, y_training, X_testing, y_testing)
n_est=x[1]
max_feat =x[2]
max_depth = x[3]
max_Sam_Split= x[4]
rfc = RFClassifier(n_est, max_feat, max_depth, max_sam_split )
rfc.fit(X_training, y_training)
y_prediction = rf_classifier.predict(X_testing)
f1 = f1_score(y_testing, y_prediction)
fit=1-f1
return fit
```

The next steps are a checking of acceptance of new solution, increasing r_i and reducing A_i . If ($rand > A_i$ and $f(x_i) < \mathcal{F}(x_*)$) then the current solution needs to be updated using the solutions found by using (5) and (6), r_i is increased and A_i is reduced.

$$A_i^{t+1} = \alpha A_i^t, \quad (6)$$

$$r_i^{t+1} = r_i^0 [1 - \exp(-\gamma t)], \quad (7)$$

The range of α is $0 < \alpha < 1$, whereas $\gamma > 0$. Its value is α , and γ could be adjusted with $\alpha=\gamma=0.9$ to facilitate the search process. According to Yang [39], the procedure for searching could be made simpler by using A_i and r_i to denote the values of 1 and 1, with $\alpha=\gamma=0.9$. The bats are sorted in the final stage to obtain the best solution(x_*). If one of the termination conditions, the maximum number of iterations or the fitness improvement, is satisfied, the proposed method's program will be terminated. A presumption is made. When the global best's fitness does not increase after 20 rounds, the global optimal point has been identified. After BA determines the RF parameters, the classification model is utilized. The looping process will be stopped

and the program's output will be saved if the termination condition is satisfied. The RF model with the appropriate parameters is the result of this process. Evaluation of classification models for DM diagnosis using RF-BA on training data as shown in Algorithm 1 and evaluation of classification models for DM diagnosis using RF-BA on testing data as shown in Algorithm 2.

Algorithm 1. Evaluation of a classification model for DM diagnosis using RF-BA on training data.

Input: the training data (X_{train}) with size of $n \times m$, hyperparameters of RF, y_{train} (Training data set's class label)

Output: accuracy, recall, precision, f_1 score.

1. Train RF Model using training data.
2. Compute the label prediction y_{pred} based on RF-BA.
3. Compute Accuracy, Recall, Precision and f_1 Score.

Algorithm 2. Evaluation of a classification model for DM diagnosis using RF-BA on testing data.

Input: the testing data, RF model, $y_{testing}$ (each testing data's class labels)

Output: accuracy, recall, precision, f_1 score.

1. Compute the label prediction y_{pred} based on RF-BA.
2. Compute Accuracy, Recall, Precision and f_1 Score.

2.4. Setting the parameters of the proposed method

The following lists the range of RF parameters that are permitted in this study: The maximum depth is [1,10], the lowest sample split is [1,20], the minimum sample split leaf is [1,20], and the number of estimators is [10,100]. These parameters are derived from earlier studies [20]. Bat loudness (A), maximum iteration (t_{max}), number of bats ($Nbat$), pulse rate (r_0), α , γ , maximum frequency (f_{max}), and minimum frequency (f_{min}) are among the BA characteristics that are employed. The BA used's parameter settings are as follows: $f_{min} = 0$, $f_{max} = 2$, $Nbat = 100$, $t_{max} = 500$, $A = 1$, $r_0 = 1$, $\alpha = 0.97$, and $\gamma = 0.1$.

2.5. Evaluating the proposed method

The metrics assessment must be computed to evaluate the classification model. The RF-BA hyperparameters optimization is used to build the classification model. Consequently, the model is assessed using the testing data. The accuracy, recall, precision, and f_1 score of the training and testing data are calculated to evaluate the classification model's efficacy. The following is a description of each evaluation metric.

- In (8) is used to calculate the accuracy.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

Positive data that is correctly interpreted is referred to as a true positive (TP). True negative (TN) is the quantity of tuples correctly classified as negative. The amount of erroneously identified tuples in the negative class is known as false positives (FP). False negative (FN) refers to the quantity of tuples that are classified as negative. The ratio of accurate predictions to total tuples is known as accuracy.

- In (9) indicates recall that is used to measure the proportion of correctly recognized positive patterns.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (9)$$

- The definition of precision is given by (10). It is computed as the ratio of all the tuples in the positive category to the correctly predicted positive class.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (10)$$

- In (11) is utilized to compute the f_1 score. It is computed by using the harmonic mean of recall and precision.

$$f_1 \text{ score} = 2 \times \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (11)$$

3. RESULTS AND DISCUSSION

The accuracy, precision, recall, and f_1 score are used to assess the proposed approach. The dataset should be examined for missing values as the initial step. The findings indicate that there are no missing values in the dataset. The min-max normalization approach is then used to convert the dataset. All of the data has the same range [0,1], according to the normalization results. The goal of this process is to make the RF approach more effective. As a result, the dataset is separated into training and testing data subsets. The percentages of testing and training data are 30 and 70%, accordingly Tables 1 to 4 show the

effectiveness of the RF-PSO-based and RF-BA-based DM diagnosis techniques. The average performance metrics of RF, PSO-RF, and BA-RF to diagnose DM for data training and testing are displayed in Tables 1 and 3, respectively. The experiment uses several the number of bats which are 5, 10, 20, and 50 bats. RF-BA 5 means that BA uses 5 bats, and RF-BA 5 means that BA uses 10 bats. The RF-BA produced results for accuracy, precision, recall, and f_1 score of 0.7861, 0.7668, 0.8315, and 0.7978, respectively, whereas the RF-PSO produced results for accuracy, precision, recall, and f_1 score of 0.7516, 0.7346, 0.7738, and 0.7536. The traditional RF for training data yielded the following results accuracy, precision, recall, and f_1 score: 0.9953, 0.9942, 0.9964, and 0.9952, respectively.

Table 1. The performance metrics average for DM diagnosis using RF, RF-PSO, and RF-BA (training data)

Methods	t	Accuracy	Precision	Recall	f_1 Score	Computational time	Fitness
RF-BA 5	50	0.7894	0.7688	0.8370	0.8014	16.2429	0.2275
RF-BA 10	50	0.7807	0.7611	0.8275	0.7929	16.2546	0.2265
RF-BA 20	60	0.7879	0.7699	0.8298	0.7987	20.0742	0.2277
RF-BA 50	50	0.7865	0.7672	0.8317	0.7981	15.8187	0.2268
RF-BA	52.5	0.7861	0.7668	0.8315	0.7978	17.0976	0.2271
RF-PSO 5	28.12	0.7521	0.7356	0.7719	0.7533	49.0971	0.2259
RF-PSO 10	46.44	0.7475	0.7314	0.7664	0.7485	179.6955	0.2277
RF-PSO 20	41.36	0.7567	0.7383	0.7859	0.7613	918.0722	0.2315
RF-PSO 50	41.52	0.7501	0.7332	0.7709	0.7515	640.1197	0.2297
RF-PSO	39.36	0.7516	0.7346	0.7738	0.7536	446.7461	0.22869
RF	-	0.9953	0.9942	0.9964	0.9952	5.9194	-

Table 2. The performance metrics standard deviation of RF, RF-PSO and RF-BA for DM diagnosis (training data)

Methods	t	Accuracy	Precision	Recall	f_1 Score	Computational time	Fitness
RF-BA 5	0.0000	0.0396	0.0384	0.0356	0.0365	6.1681	0.0118
RF-BA 10	0.0000	0.0185	0.0189	0.0169	0.0167	5.5662	0.0081
RF-BA 20	0.0000	0.0253	0.0235	0.0256	0.0240	8.1078	0.0098
RF-BA 50	0.0000	0.0324	0.0319	0.0279	0.0295	4.0908	0.0092
RF-BA	0	0.02895	0.02818	0.0265	0.02668	5.98323	0.0097
RF-PSO 5	7.6829	0.0074	0.0068	0.0100	0.0079	23.4423	0.0098
RF-PSO 10	11.2475	0.0087	0.0069	0.0140	0.0098	80.6586	0.0102
RF-PSO 20	10.0825	0.0120	0.0101	0.0234	0.0156	905.5180	0.0099
RF-PSO 50	10.6031	0.0151	0.0129	0.0201	0.0158	310.5717	0.0144
RF-PSO	9.904	0.0108	0.009175	0.016875	0.012275	330.0476	0.0110
RF	0	1.744×10^{-5}	0.000184	0.00018	1.735×10^{-5}	0.11033	-

Table 3. The performance metrics average of RF, RF-PSO and RF-BA for DM diagnosis (testing data)

Methods	Accuracy	Precision	Recall	f_1 score
RF-BA 5	0.7686	0.7688	0.7796	0.7740
RF-BA 10	0.7666	0.7685	0.7742	0.7713
RF-BA 20	0.7688	0.7717	0.7747	0.7730
RF-BA 50	0.7679	0.7724	0.7707	0.7715
RF-BA	0.7680	0.7703	0.7748	0.7725
RF-PSO 5	0.7695	0.7718	0.7683	0.7717
RF-PSO 10	0.7689	0.7711	0.7664	0.7710
RF-PSO 20	0.7642	0.7668	0.7715	0.7663
RF-PSO 50	0.7708	0.7731	0.7698	0.7730
RF-PSO	0.7684	0.7707	0.7690	0.7705
RF	0.7371	0.7181	0.7783	0.7470

Table 4. The performance metrics standard deviation of RF, RF-PSO, and RF-BA for DM diagnosis (testing data)

Methods	Accuracy	Precision	Recall	f_1 score
RF-BA 5	0.0089	0.0106	0.01349	0.0087
RF-BA 10	0.0070	0.0090	0.01311	0.0073
RF-BA 20	0.0076	0.0089	0.01556	0.0085
RF-BA 50	0.0085	0.0069	0.01600	0.0098
RF-BA	0.0080	0.0089	0.01454	0.0086
RF-PSO 5	0.0100	0.0113	0.01542	0.0102
RF-PSO 10	0.0083	0.0080	0.01154	0.0085
RF-PSO 20	0.0070	0.0071	0.01266	0.0077
RF-PSO 50	0.0152	0.0261	0.02838	0.0162
RF-PSO	0.0101	0.01314	0.01700	0.0106
RF	0.0014	0.00156	0.00198	0.00134

For testing data, the accuracy, precision, recall, and f_1 score obtained from the RF-BA are 0.7680, 0.7703, 0.7748, and 0.7725, respectively. Similarly, the accuracy, precision, recall and f_1 score obtained from the RF-PSO are 0.76841, 0.7707, 0.7690, and 0.7705, respectively. Using the traditional RF for testing data, the corresponding accuracy, precision, recall, and f_1 score are 0.7372, 0.7181, 0.7783, and 0.7470. As demonstrated by the results, RF-BA and RF-PSO outperform conventional RF and can resolve the overfitting issue of RF. Regarding accuracy, precision, recall, and f_1 score, RF-BA and RF-PSO do not differ much in performance.

The experimental findings also demonstrate that the PSO and BA function identically regardless of the number of particles or bats which is 5, 10, 20, and 50. Nevertheless, RF-BA's computation time is substantially quicker than RF-PSO's. The grid search approach takes longer to compute than the RF-PSO and RF-BA methods. The RF-PSO is to blame, and the RF-BA does not have to experiment with every possible combination of RF hyperparameters. The fitness values that RF-PSO and RF-BA produce are very similar. Tables 2 and 4 demonstrate that RF-PSO and RF-BA produced good variances for all bat/particle counts. Consequently, five particles or bats is the suggested quantity. Generally, RF using the SI technique (BA and PSO) performs far better than traditional RF. To increase performance, the RF-BA method of diagnosing DM must still be used. Optimizing the data preprocessing stages, including feature selection, can help achieve the improvement. Process optimization on BA, such as a robust population initiation to increase the global optima search, can also be used to improve.

4. CONCLUSION

The conclusion drawn from the examination of the experimental results is the RF-BA is better than the traditional RF and the RF-PSO for the DM diagnosis. The over-fitting situations on the conventional RF for diagnosing DM can be avoided based on the RF-BA and RF-PSO. Compared to RF-PSO, RF-BA has a faster computation time. Compared to the grid search approach, the RF-PSO and RF-BA also yield shorter computation times. For every bat or particle count, both RF-PSO and RF-BA produced good variances. Consequently, it is advised that there be five particles in each bat. Even though RF-BA takes less time to compute than RF-PSO, a faster BA procedure is still required. In addition, the RF-BA method of diagnosing DM is still needed. Optimizing the data preprocessing stages, including feature selection, can help it get better. Additionally, the BA process still has to be enhanced in order to increase the worldwide search for optimal results. The suggested approach has a considerable potential of being used in the future to assist individuals with early diabetes diagnosis in a fast, low-cost, and highly accurate manner.

ACKNOWLEDGEMENTS

All authors would like to say their gratitude to the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia's, DIKTI for support the research through Cooperative Research Grant - Domestic 2023 (*Hibah Penelitian Kerjasama - Dalam Negeri 2023*), under contract number 199/E5/PG.020.00.PL/2023.

REFERENCES

- [1] S. Fattaheian-Dehkordi, R. Hojjatifard, M. Saeedi, and M. Khanavi, "A review on antidiabetic activity of centaurea spp.: a new approach for developing herbal remedies," *Evidence-Based Complementary and Alternative Medicine*, vol. 2021, pp. 1–23, 2021, doi: 10.1155/2021/5587938.
- [2] A. M. Hutaape and C. Susanto, "Hypoglycemic potential of Aloe vera in diabetes mellitus induced by diabetogenic substances and high fat diet: A systematic meta-analysis review," *International Journal of Applied Dental Sciences*, vol. 7, no. 3, pp. 360–368, 2021, doi: 10.22271/oral.2021.v7.i3f.1322.
- [3] K. Papatheodorou, M. Banach, E. Bekiari, M. Rizzo, and M. Edmonds, "Complications of diabetes 2017," *Journal of Diabetes Research*, vol. 2018, pp. 1–4, 2018, doi: 10.1155/2018/3086167.
- [4] D. Tomic, J. E. Shaw, and D. J. Magliano, "The burden and risks of emerging complications of diabetes mellitus," *Nature Reviews Endocrinology*, vol. 18, no. 9, pp. 525–539, 2022, doi: 10.1038/s41574-022-00690-7.
- [5] K. J. Brahmbhatt *et al.*, "Association of mean platelet volume with vascular complications in the patients with type 2 diabetes mellitus," *Cureus*, vol. 14, no. 9, 2022, doi: 10.7759/cureus.29316.
- [6] J. C. Hartz, S. D. Ferranti, and S. Gidding, "Hypertriglyceridemia in diabetes mellitus: implications for pediatric care," *Journal of the Endocrine Society*, vol. 2, no. 6, pp. 497–512, 2018, doi: 10.1210/jse.2018-00079.
- [7] H. Aryan, A. Najmaldin, and A. Gohari, "Mortality rate and related risk factors in hospitalized coronavirus disease 2019 patients with diabetes: a single-center study," *Galen Medical Journal*, vol. 11, 2022, doi: 10.31661/gmj.v11i.2590.
- [8] A. D. Deshpande, M. Harris-Hayes, and M. Schootman, "Epidemiology of diabetes and diabetes-related complications," *Physical Therapy*, vol. 88, no. 11, pp. 1254–1264, 2008, doi: 10.2522/ptj.20080020.
- [9] M. Grujicic, A. Salapura, G. Jovanovic, A. Figurek, D. Zrnic, and A. Grbic, "Non-diabetic kidney disease in patients with type 2 diabetes mellitus-11-year experience from a single center," *Medical Archives*, vol. 73, no. 2, pp. 87–91, 2019, doi: 10.5455/medarh.2019.73.87-91.
- [10] S. Kamle, M. Holay, P. Patil, and P. Tayde, "Clinical profile and outcome of diabetic ketoacidosis in type 1 and type 2 diabetes: a comparative study," *Vidarbha Journal of Internal Medicine*, vol. 32, 2022, doi: 10.25259/VJIM_11_2021.
- [11] A. Armayani *et al.*, "Effect of hydrogel use on healing diabetic foot ulcers: systematic review," *Open Access Macedonian Journal of Medical Sciences*, vol. 10, no. F, pp. 448–453, 2022, doi: 10.3889/oamjms.2022.9835.
- [12] B. Hidayat, R. V. Ramadani, A. Rudijanto, P. Soewondo, K. Suastika, and J. Y. Siu Ng, "Direct medical cost of type 2 diabetes

mellitus and its associated complications in Indonesia," *Value in Health Regional Issues*, vol. 28, pp. 82–89, 2022, doi: 10.1016/j.vhri.2021.04.006.

[13] V. Kavuru, R. S. Senger, J. L. Robertson, and D. Choudhury, "Analysis of urine Raman spectra differences from patients with diabetes mellitus and renal pathologies," *PeerJ*, vol. 11, p. e14879, 2023, doi: 10.7717/peerj.14879.

[14] R. P. S. Makkar, A. Monga, A. Arora, S. Mukhopadhyay, and A. K. Gupta, "Self-referral to specialists – a dodgy proposition," *International Journal of Health Care Quality Assurance*, vol. 16, no. 2, pp. 87–89, 2003, doi: 10.1108/09526860310465591.

[15] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on MRI," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 102–127, 2019, doi: 10.1016/j.zemedi.2018.11.002.

[16] H. S. R. Rajula, G. Verlato, M. Manchia, N. Antonucci, and V. Fanos, "Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment," *Medicina*, vol. 56, no. 9, 2020, doi: 10.3390/medicina56090455.

[17] A. Mekrache, A. Bradai, E. Moulay, and S. Dawaliby, "Deep reinforcement learning techniques for vehicular networks: Recent advances and future trends towards 6G," *Vehicular Communications*, vol. 33, 2022, doi: 10.1016/j.vehcom.2021.100398.

[18] K. Tan, D. Bremner, J. L. Kernev, L. Zhang, and M. Imran, "Machine learning in vehicular networking: An overview," *Digital Communications and Networks*, vol. 8, no. 1, pp. 18–24, 2022, doi: 10.1016/j.dcan.2021.10.007.

[19] Z. Qiao *et al.*, "An enhanced Runge Kutta boosted machine learning framework for medical diagnosis," *Computers in Biology and Medicine*, vol. 160, 2023, doi: 10.1016/j.combiomed.2023.106949.

[20] X. Chen, X. Liu, Y. Wu, Z. Wang, and S. H. Wang, "Research related to the diagnosis of prostate cancer based on machine learning medical images: A review," *International Journal of Medical Informatics*, vol. 181, 2024, doi: 10.1016/j.ijmedinf.2023.105279.

[21] S. Madanian *et al.*, "Speech emotion recognition using machine learning — A systematic review," *Intelligent Systems with Applications*, vol. 20, 2023, doi: 10.1016/j.iswa.2023.200266.

[22] E. C. P. Neto, S. Dadkhah, S. Sadeghi, H. Molyneaux, and A. A. A. Ghorbani, "A review of machine learning (ML)-based IoT security in healthcare: A dataset perspective," *Computer Communications*, vol. 213, pp. 61–77, 2024, doi: 10.1016/j.comcom.2023.11.002.

[23] D. R. Wijaya, F. Afianti, A. Arifianto, D. Rahmawati, and V. S. Kodogiannis, "Ensemble machine learning approach for electronic nose signal processing," *Sensing and Bio-Sensing Research*, vol. 36, 2022, doi: 10.1016/j.sbsr.2022.100495.

[24] V. Bharilya and N. Kumar, "Machine learning for autonomous vehicle's trajectory prediction: A comprehensive survey, challenges, and future research directions," *Vehicular Communications*, vol. 46, 2024, doi: 10.1016/j.vehcom.2024.100733.

[25] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sánchez, "An assessment of the effectiveness of a random forest classifier for land-cover classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 67, pp. 93–104, 2012, doi: 10.1016/j.isprsjprs.2011.11.002.

[26] M. Shatnawi, N. Zaki, and P. D. Yoo, "Protein inter-domain linker prediction using random forest and amino acid physiochemical properties," *BMC Bioinformatics*, vol. 15, no. S8, 2014, doi: 10.1186/1471-2105-15-S16-S8.

[27] Z. Yikun, H. Yingjie, Z. Haixiao, L. Jiahao, L. Yijin, and L. Jinjun, "Classification method of voltage sag sources based on sequential trajectory feature learning algorithm," *IEEE Access*, vol. 10, pp. 38502–38510, 2022, doi: 10.1109/ACCESS.2022.3164675.

[28] N. K. Mishra *et al.*, "Automatic lesion border selection in dermoscopy images using morphology and color features," *Skin Research and Technology*, vol. 25, no. 4, pp. 544–552, 2019, doi: 10.1111/srt.12685.

[29] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, "A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector," *IEEE Access*, vol. 7, pp. 60134–60149, 2019, doi: 10.1109/ACCESS.2019.2914999.

[30] A. Muneer, R. F. Ali, A. Alghamdi, S. M. Taib, A. Almaghthawi, and E. A. A. Ghaleb, "Predicting customers churning in banking industry: A machine learning approach," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 26, no. 1, pp. 539–549, 2022, doi: 10.11591/ijeecs.v26.i1.pp539-549.

[31] Z. Zhao, W. Zhou, Z. Qiu, A. Li, and J. Wang, "Research on ctrip customer churn prediction model based on random forest," in *International Conference on Business Intelligence and Information Technology*, 2022, pp. 511–523. doi: 10.1007/978-3-030-92632-8_48.

[32] M. Pal and S. Parija, "Prediction of heart diseases using random forest," *Journal of Physics: Conference Series*, vol. 1817, no. 1, 2021, doi: 10.1088/1742-6596/1817/1/012009.

[33] N. K. Yego, J. Kasozi, and J. Nkurunziza, "A comparative analysis of machine learning models for the prediction of insurance uptake in Kenya," *Data*, vol. 6, no. 11, 2021, doi: 10.3390/data6110116.

[34] V. G. Krishnan, S. D. Raj, S. Lokesh, and S. Sudharshan, "Credit card fraud detection using random forest algorithm," *International Journal for Research in Applied Science and Engineering Technology*, vol. 7, no. 3, pp. 1199–1202, 2019, doi: 10.22214/ijraset.2019.3215.

[35] M. Minnoor and V. Baths, "Diagnosis of breast cancer using random forests," *Procedia Computer Science*, vol. 218, pp. 429–437, 2023, doi: 10.1016/j.procs.2023.01.025.

[36] N. Zhu, C. Zhu, L. Zhou, Y. Zhu, and X. Zhang, "Optimization of the random forest hyperparameters for power industrial control systems intrusion detection using an improved grid search algorithm," *Applied Sciences*, vol. 12, no. 20, 2022, doi: 10.3390/app122010456.

[37] A. K. Kordon, "Swarm intelligence: the benefits of swarms," in *Applying Computational Intelligence*, Berlin, Heidelberg: Springer, 2010, pp. 145–174. doi: 10.1007/978-3-540-69913-2_6.

[38] W. Yang, R. Li, Y. Yuan, and X. Mou, "Economic dispatch using modified bat algorithm," *Frontiers in Energy Research*, vol. 10, 2022, doi: 10.3389/fenrg.2022.977883.

[39] X.-S. Yang, "A new metaheuristic bat-inspired algorithm," *arXiv-Mathematics*, pp. 1–10, 2010.

[40] S. Anam, Z. Fitriah, N. Hidayat, and M. H. A. A. Maulana, "Classification model for diabetes mellitus diagnosis based on k-means clustering algorithm optimized with bat algorithm," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 1, 2023, doi: 10.14569/IJACSA.2023.0140172.

[41] N. Vani and D. Vinod, "A comparative analysis on random forest algorithm over k-means for identifying the brain tumor anomalies using novel CT scan with MRI scan," in *2022 International Conference on Business Analytics for Technology and Security (ICBATS)*, 2022, pp. 1–6. doi: 10.1109/ICBATS54253.2022.9759036.

[42] S. Srivastava and S. K. Sahana, "Application of bat algorithm for transport network design problem," *Applied Computational Intelligence and Soft Computing*, vol. 2019, pp. 1–12, 2019, doi: 10.1155/2019/9864090.

[43] A. Asokan, D. E. Popescu, J. Anitha, and D. J. Hemanth, "Bat algorithm based non-linear contrast stretching for satellite image enhancement," *Geosciences*, vol. 10, no. 2, 2020, doi: 10.3390/geosciences10020078.

[44] T. A. Rashid *et al.*, "An improved BAT algorithm for solving job scheduling problems in hotels and restaurants," in *Studies in Computational Intelligence*, Springer, Cham, 2021, pp. 155–171. doi: 10.1007/978-3-03-72711-6_9.

[45] S. Anam and Z. Fitriah, "Early blight disease segmentation on tomato plant using k-means algorithm with swarm intelligence-based algorithm," *International Journal of Mathematics and Computer Science*, vol. 16, no. 4, pp. 1217–1228, 2021.

[46] L. Breiman, "Random forests," in *Machine Learning*, Springer, 2001, pp. 5–32, doi: 10.1023/A:1010933404324.

BIOGRAPHIES OF AUTHORS



Syaiful Anam received a Doctor of Natural Science and Mathematics degree from Yamaguchi University, Japan in 2015. He also received his Bachelor Degree in Mathematics from Brawijaya University, Indonesia in 2001 and his Master Degree from Sepuluh Nopember Institute of Technology, Indonesia in 2006. He is currently an Assistant Professor at Department of Mathematics, Brawijaya University, Malang, Indonesia. His research includes data science, computational intelligence, machine learning, digital image processing, and computer vision. He has published over 35 papers in international journals and conferences. He can be contacted at email: syaiful@ub.ac.id.



Fidia Deny Tisna Amijaya holds the Master Degree in Mathematics from the Brawijaya University, Malang, Indonesia, with the master thesis “Hybrid greedy algorithm - particle swarm optimization - genetic algorithm (Hybrid GPSOGA)”. He also received his Bachelor Degree in Mathematics from Brawijaya University, Indonesia in 2011. He is an Assistant Professor in Department of Mathematics, Faculty of Mathematics and Natural Sciences, Mulawarman University, Samarinda, Indonesia. His research interests are in applied mathematics, data mining, and computational intelligence. He can be contacted at email: fidia@fmipa.unmul.ac.id.



Satrio Hadi Wijoyo holds the Master Degree in Informatics Engineering from the Sepuluh Nopember Institute of Technology, Surabaya, Indonesia. He also received his Bachelor Degree in Mathematics from Brawijaya University, Indonesia in 2013. He is an Assistant Professor in Department of Information System, Faculty of Computer Science, Malang, Indonesia. His research interests are Education, learning, evaluation, learning media, intelligent computing, and data & information management. He can be contacted at email: satriohadi@ub.ac.id.



Dian Eka Ratnawati holds a Bachelor of Science in Mathematics, Master in Informatics Engineering and Doctor in Mathematics. She is currently lecturing with the Department of Informatics Engineering at Faculty of Science, Brawijaya University, Malang, Indonesia. Her research areas of interest include computational intelligence. She can be contacted at email: dian_ilkom@ub.ac.id.



Cynthia Ayu Dwi Lestari is a student of the Master Degree in Mathematics Brawijaya University, Malang, Indonesia. She also received his Bachelor Degree in Mathematics from Brawijaya University, Indonesia in 2019. Her research interests are computational intelligence and data science. She can be contacted at email: cynthia@student.ub.ac.id.



Muhammin Ilyas is a student of the Master Degree in Mathematics Brawijaya University, Malang, Indonesia. He also received his Bachelor Degree in Mathematics from Brawijaya University, Indonesia in 2023. His research interests are computational intelligence and data science. He can be contacted at email: chaimin@student.ub.ac.id.